Sequoia Architectural Requirements

The Salishan Conference on High Speed Computing

24 April 2008

Matt L. Leininger
Mark K. Seager

Lawrence Livermore National Laboratory

Prepared by LLNL under Contract DE-AC52-07NA27344

LLNL-PRES-403472

INFORMATION IN THESE SLIDES IS SUBJECT TO CHANGE!

Predictive Simulations Drive Petascale & Exascale Challenges

- Sequoia Mission Drivers
- Sequoia Target Architectures
- Sequoia Procurement Strategy
- Sequoia Benchmarks
- Challenges to fielding Petascale Simulation Environments

By leveraging industry trends, Sequoia will successfully deliver a petascale Uncertainty Quantification engine for the stockpile

- Sequoia Production Platform Programmatic Drivers
  - UQ Engine(2D & 3D) for LEP and RRW in the 2011-2015 timeframe
  - Predictive simulation for boost
  - Petascale integrated weapons simulations

- Programmatic drivers require unprecedented leap forward in computing power

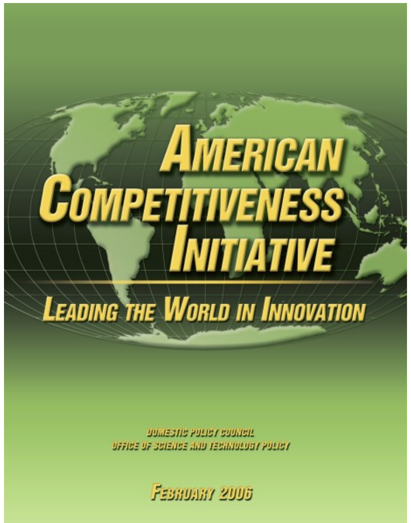
- Program needs both *Capability* and *Capacity*
  - 25-50x BGL (367TF/s) for science codes (knob removal)
  - 12-24x Purple for IDC capability runs on Purple (8,192 MPI tasks UQ Engine)

- These requirements coupled with current industry trends drive us to a different target architecture than Purple or BGL

The scientific method has fundamentally changed for the first time since Galileo

“the intersection of computer science and the sciences … has the potential to have a profound impact on science. It is a leap from the application of computing … to the *integration of computer science concepts, tools, and theorems* into the very fabric of science.”

Science 2020 Report, March 2006



AMERICAN COMPETITIVENESS INITIATIVE

LEADING THE WORLD IN INNOVATION

DOMESTIC POLICY COUNCIL
OFFICE OF SCIENCE AND TECHNOLOGY POLICY

FEBRUARY 2006

2020 COMPUTING SPECIAL: CHARTING THE INFORMATION EXPLOSION

23 March 2006 | www.nature.com/nature | $10

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

nature

LIFE ON EARTH
Methane-producing microbes arrived early

PLANT SPECIES
Do we still need them?

AVIAN FLU VIRUS
A molecular barrier to human-to-human transmission

2020 VISION

How computers will change the face of science

INSIDE
Ion channels

nature insight

NATUREJOBS
The Midwest

$10.00US $12.99CAN

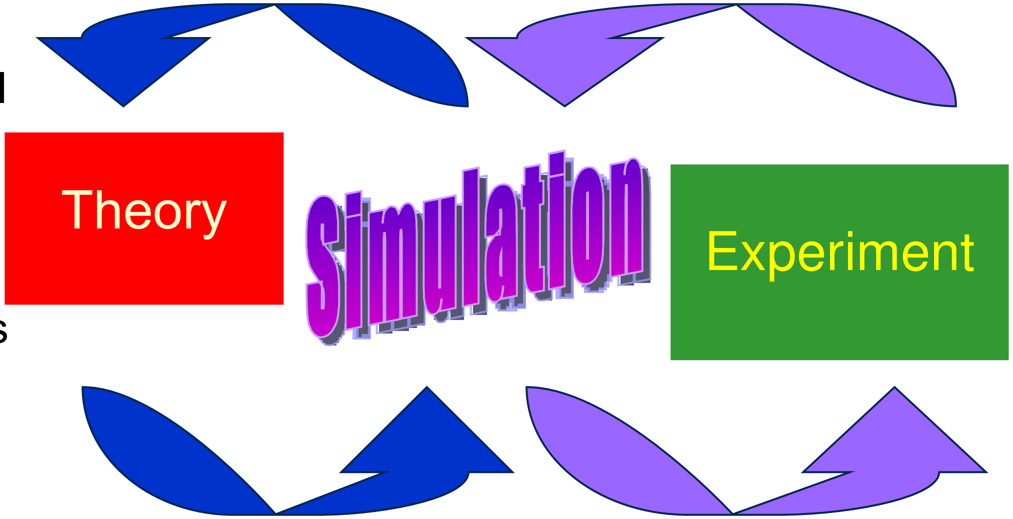
Simulation has become the critical integrating element between theory and experiment

Predictive simulation ENABLES

- Detailed predictive assessment of complex models for overarching physical problems
- Design of experiments
- Impact assessment of policy choices
- Elimination of costly physical prototypes

Predictive simulation REQUIRES

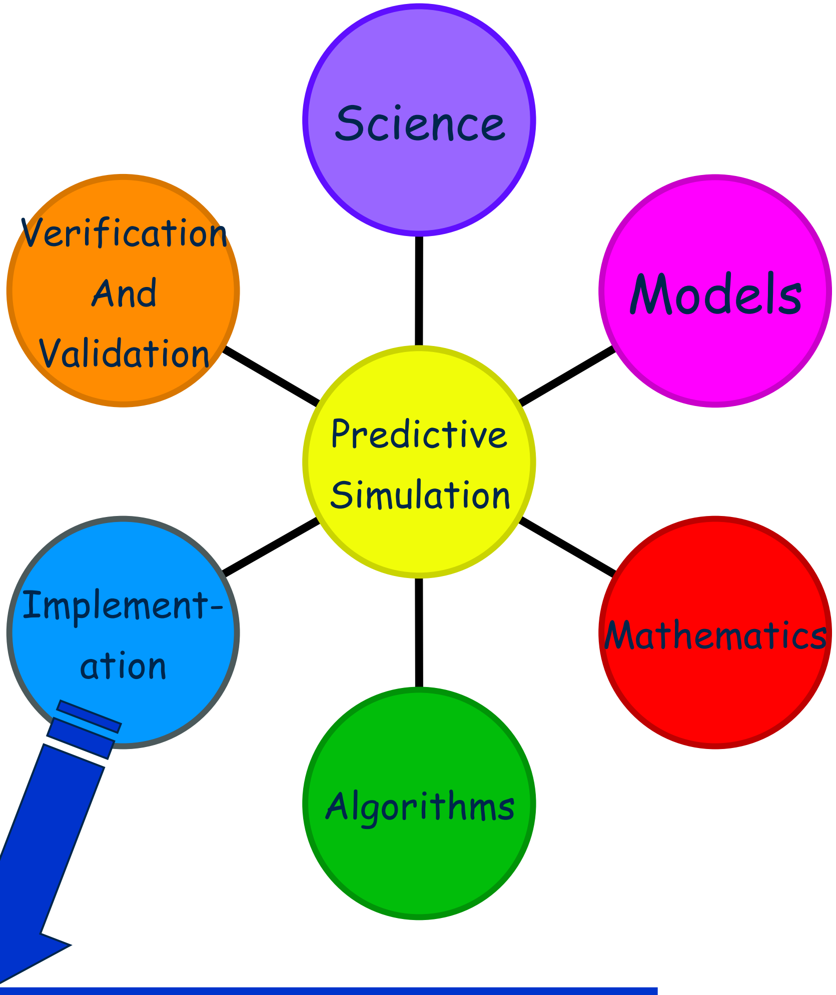
- Verification and validation of complex models (experiment)
- Development of science based models (theory)
- Databases of physical properties and catalogues of scientific data
- Petascale simulation environments



Revolution in the making: BlueGene/L at LLNL

Predictive simulation requires advances on multiple simultaneous fronts

- Increased accuracy comes from a complex interaction of
  - Improved understanding of the science at multiple scales
  - More detailed models of the underlying interacting science
  - Better mathematical treatment of the model
  - Increased accuracy and scalability of the solution algorithms
  - Code development for faster systems
  - Verification and Validation
    - •Are the equations solved correctly?
    - •Are the right equations being solved?



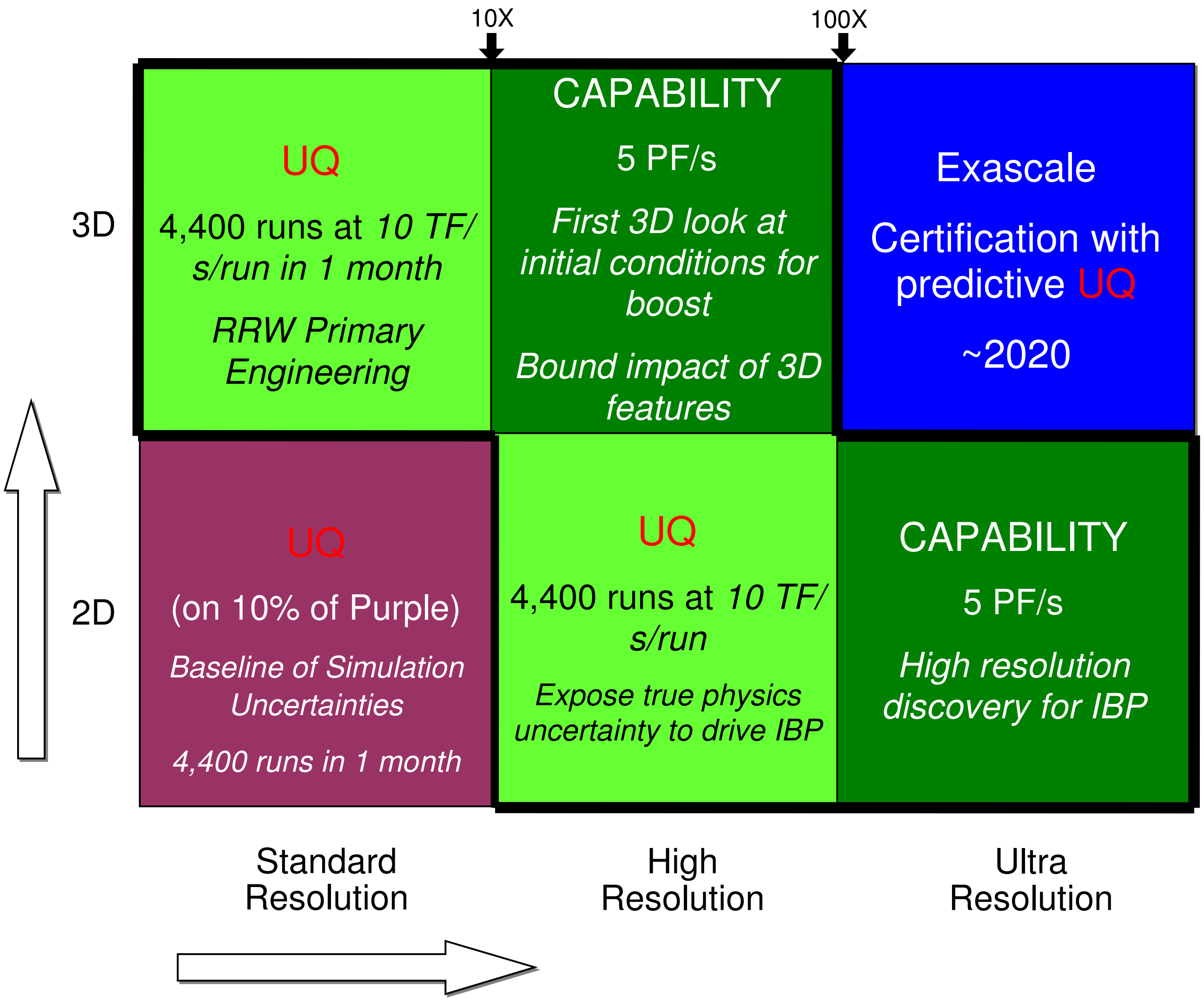
If your petascale applications strategy is "port and scale the codes," then guess again. This is only as small part of the overall challenge!

Predicting stockpile performance drives five separate classes of petascale calculations

1. Quantifying uncertainty (for all classes of simulation)
2. Identify and model missing physics (e.g., boost)
3. Improving accuracy in material property data
4. Improving models for known physical processes
5. Improving the performance of complex models and algorithms in macro-scale simulation codes

Each of these mission drivers require petascale computing

Sequoia is the key integrating tool for Stockpile Stewardship in 2011-2015 time frame



	Standard Resolution	High Resolution	Ultra Resolution
3D	UQ 4,400 runs at *10 TF/s/run in 1 month* *RRW Primary Engineering*	CAPABILITY 5 PF/s *First 3D look at initial conditions for boost* *Bound impact of 3D features*	Exascale Certification with predictive UQ ~2020
2D	UQ (on 10% of Purple) *Baseline of Simulation Uncertainties* *4,400 runs in 1 month*	UQ 4,400 runs at *10 TF/s/run* *Expose true physics uncertainty to drive IBP*	CAPABILITY 5 PF/s *High resolution discovery for IBP*

Sequoia will provide credible UQ for stockpile certification

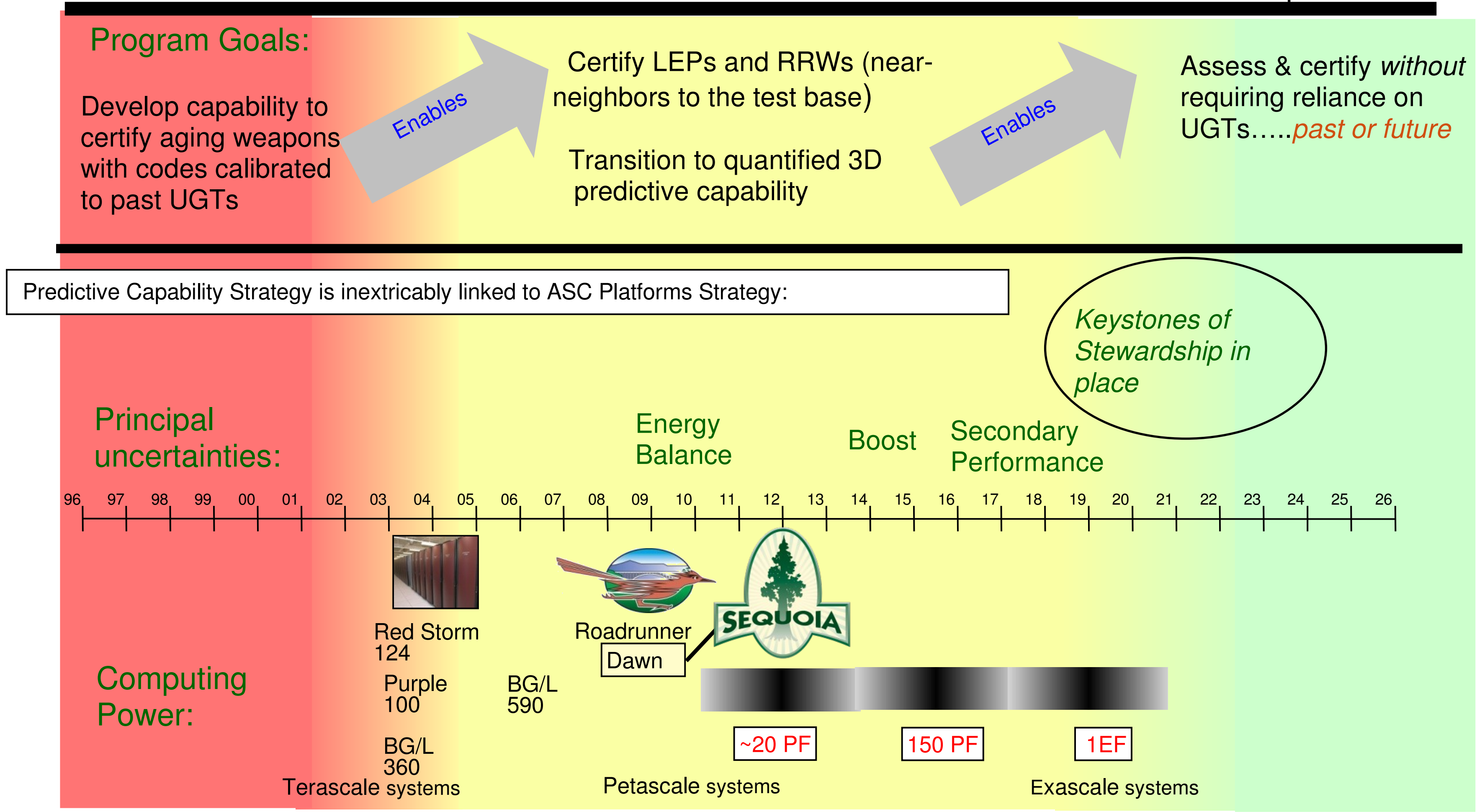
- Execution of Integrated Boost Program (IBP) example of weapons science that requires Sequoia – IBP can't be done on Purple
- IBP is necessary to achieve certification with predictive UQ

ASC Continues with roadmap to exascale

Past

2008-2018

Transformed Complex



Program Goals:

Develop capability to certify aging weapons with codes calibrated to past UGTs

Enables

Certify LEPs and RRWs (near-neighbors to the test base)

Transition to quantified 3D predictive capability

Enables

Assess & certify *without* requiring reliance on UGTs…..*past or future*

Predictive Capability Strategy is inextricably linked to ASC Platforms Strategy:

Keystones of Stewardship in place

Principal uncertainties:

Energy Balance

Boost

Secondary Performance

Computing Power:

Red Storm 124

Purple 100

BG/L 360

BG/L 590

Roadrunner

Dawn

~20 PF

150 PF

1EF

Terascale systems

Petascale systems

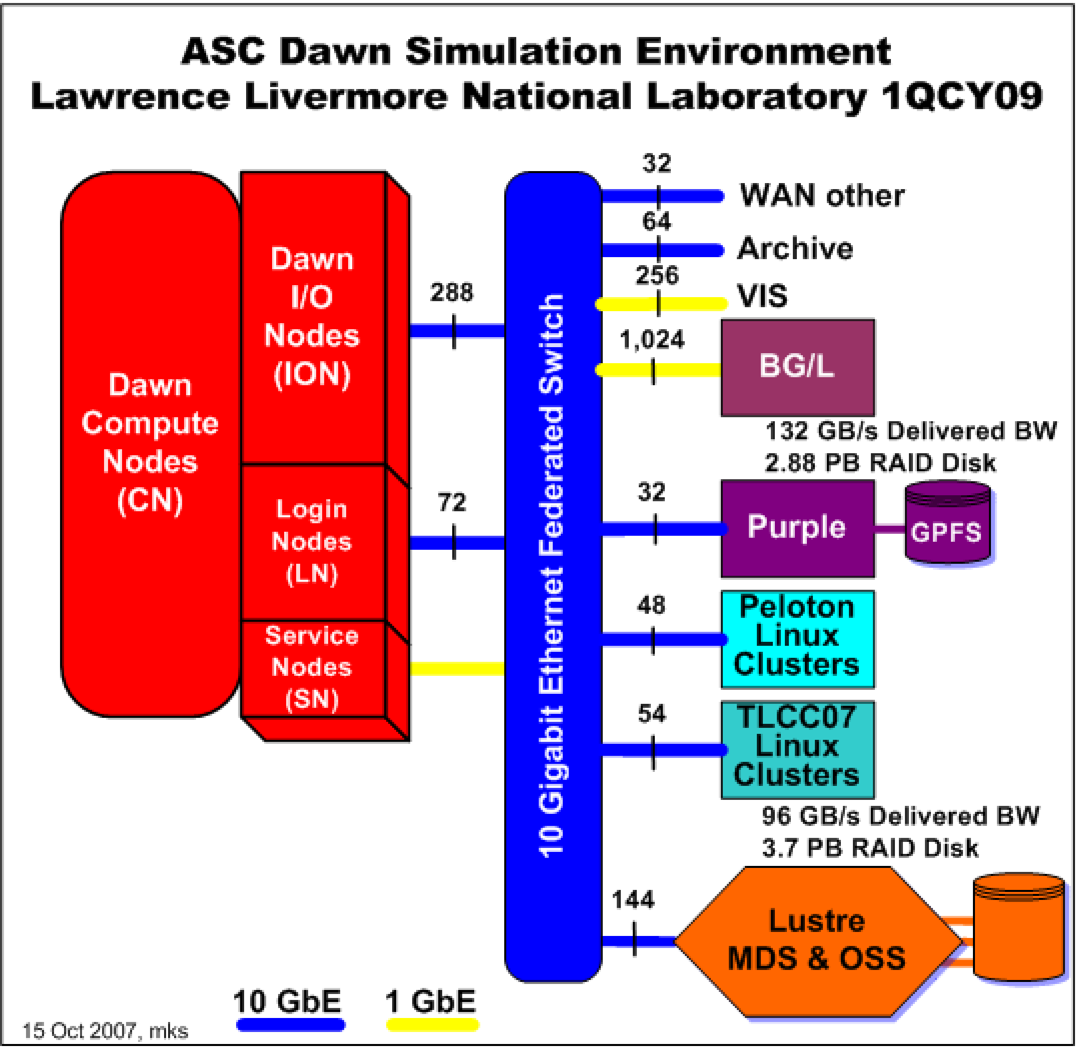
Exascale systems

ASC RoadRunner and Sequoia is the dawn of the petascale era for predictive weapons science

Sequoia Procurement Strategy Draws on ASC Experience with Successfully Delivering Five Generations of Platforms to Stockpile Stewardship Program

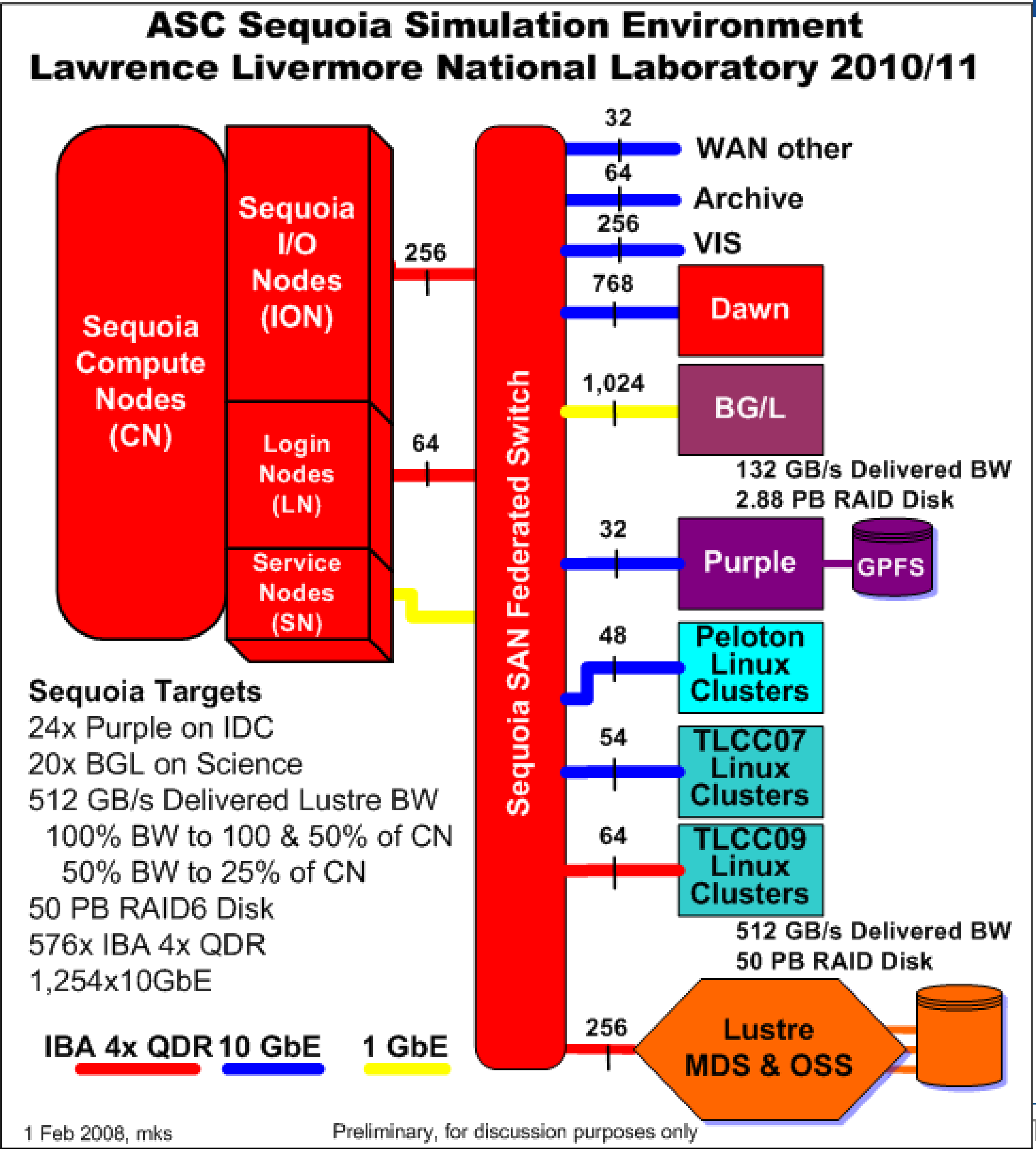
- **Two Major Deliverables**
  - Petascale Scaling "Dawn" Platform in 2008
  - Petascale Weapons "Sequoia" Platform in 2011
- **Lessons learned from previous capability and capacity procurements**
  - Leverage best-of-breed for platform, file system, SAN and storage
  - Major Sequoia procurement is for long term platform partnership
  - Three R&D partnerships to incentivize bidders to stretch goals
  - Risk reduction built into overall strategy from day-one
- **Drive procurement with single peak mandatory**
  - Target Peak+Sustained on marquee benchmarks
  - Timescale, budget, technical details as target requirements
  - Include TCO factors such as power

To Minimize Risk, Dawn Deployment Extends the Existing Purple and BG/L Integrated Simulation Environment



- ASC Dawn is the initial delivery system for Sequoia
- Code development platform and scaling for Sequoia
- 0.5 petaFLOP/s peak for ASC production usage
- Target production 2009-2014
- Dawn Component Scaling
  - Memory B:F = 0.3
  - Mem BW B:F = 1.0
  - Link BW B:F = 0.1
  - Min Bisect B:F = 0.001
  - SAN GB/s:PF/s = 384
  - F is peak FLOP/s

Sequoia Target Architecture in Integrated Simulation Environment Enables a Diverse Production Workload



ASC Sequoia Simulation Environment
Lawrence Livermore National Laboratory 2010/11

Sequoia Targets
24x Purple on IDC
20x BGL on Science
512 GB/s Delivered Lustre BW
  100% BW to 100 & 50% of CN
    50% BW to 25% of CN
50 PB RAID6 Disk
576x IBA 4x QDR
1,254x10GbE

IBA 4x QDR 🟥 **10 GbE** 🟦 **1 GbE** 🟨

1 Feb 2008, mks

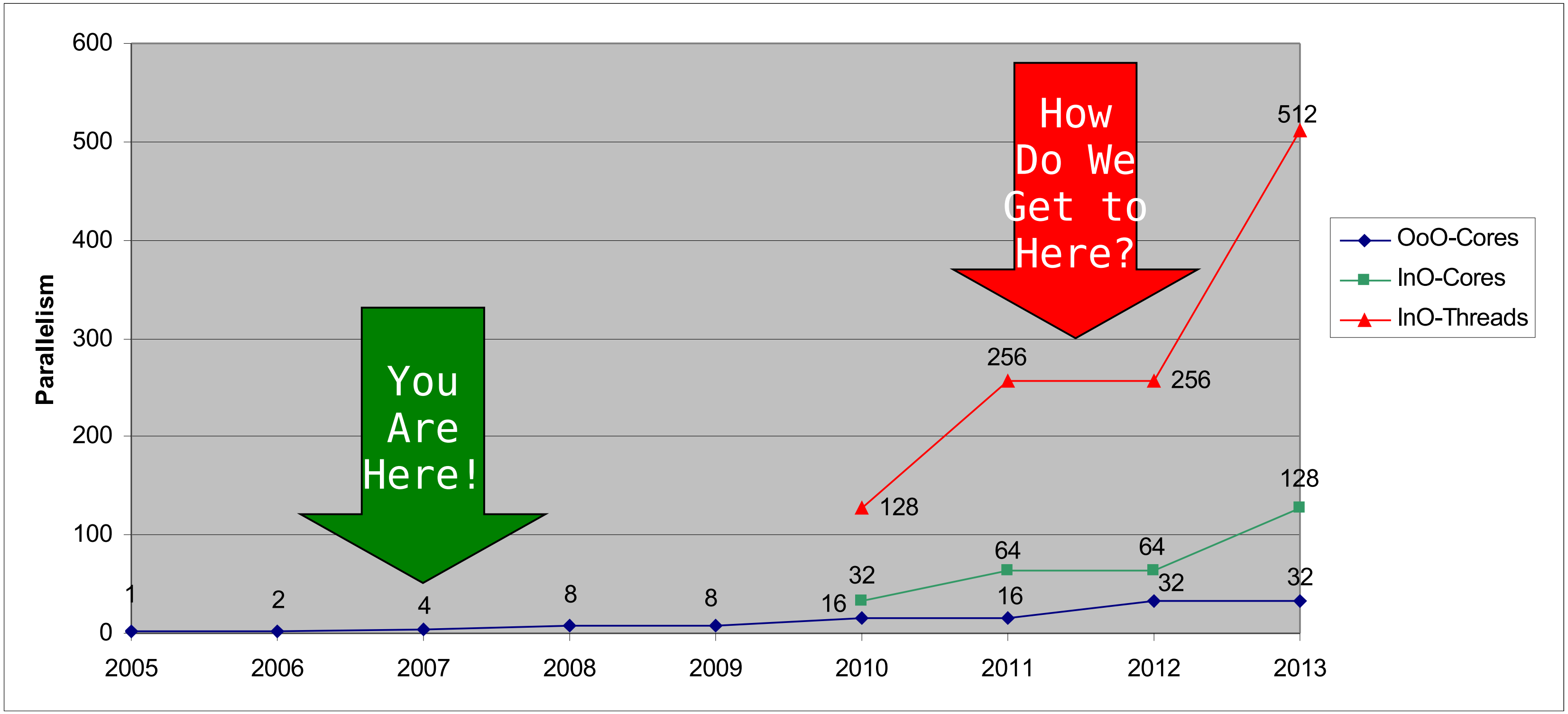
Preliminary, for discussion purposes only

- Diverse usage models drive platform and simulation environment requirements
  - Will be 2D ultra-res and 3D high-res Quantification of Uncertainty engine
  - 3D Science capability for known unknowns and unknown unknowns
- Peak of 14 petaFLOP/s with option for 20 petaFLOP/s
- Target production 2011-2016
- Sequoia Component Scaling
  - Memory B:F = 0.08
  - Mem BW B:F = 0.2
  - Link BW B:F = 0.15
  - Min Bisect B:F = 0.003
  - SAN BW GB/:PF/s = 25.6
  - F is peak FLOP/s

Livermore’s role integrates “best of breed” procurements to deliver highly usable petascale simulation environment

- The thing called “*Sequoia procurement or Sequoia RFP*” is for the platforms
  - Platform R&D partnerships solicited with Sequoia RFP
  - Identify “gaps” between Sequoia requirements and vendor offerings
  - One for winner, one for runner-up
- Lustre storage procurement for long term partnership and both generations of platforms
  - Storage R&D partnership solicited with Sequoia Storage RFP
- Storage Area Network procurement for long term partnership and both generations of platforms
- Separate test and evaluation resource for Sequoia
  - Unique long term partnership with technology providers
  - This testing model enabled world’s first integrated simulation environment for BG/L

How many cores are you coding for?



Microprocessor parallelism will increase exponentially in the next decade

From Herb Sutter <hsutter@microsoft.com>

How much parallelism will be required to sustain petaFLOP/s in 2011?

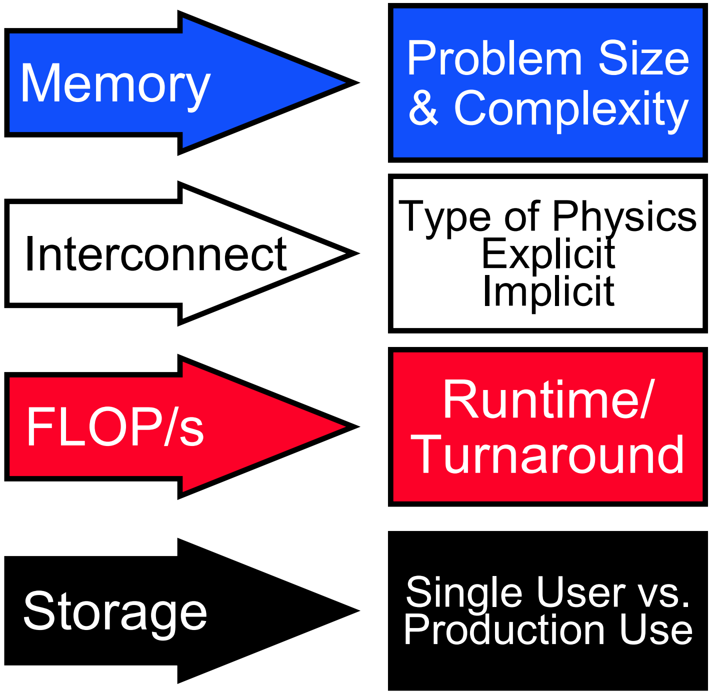
- Hypothetical low power machines will feature 1.6M to 6.6M way parallelism
  - 32-64 cores per processor and up to 2-4 threads per core
  - Assume 1 socket nodes and 25.6K nodes
- Hypothetical Intel terascale chip petascale system yields 1.5M way parallelism
  - 80 cores per processor
  - Assume 4 socket nodes and 4,608 nodes (32 SU of 144 nodes with IBA)
- Holy cow, this is about 12-48x BlueGene/L!

Multicore processors have non-intuitive impact on other machine characteristics

- Memory is the most critical machine characteristic
- ASC applications require >1GiB/MPI task
- If we map MPI tasks directly to cores
  - 64GiB/node on Low Power ➔ 1.6PiB of memory and that is 4x too expensive, if we could build and power it
    - This drives us to think in terms of fewer MPI tasks/node
  - 320GiB/node on Intel ➔ 1.5PiB of memory is also a problem

System Component

Effect on SSMP† Program



Memory

Problem Size & Complexity

Interconnect

Type of Physics
Explicit
Implicit

FLOP/s

Runtime/ Turnaround

Storage

Single User vs. Production Use

†From Seager Dec 1997 platforms talk

How much parallelism will be required to sustain petaFLOP/s in 2011?

- How do applications sustain a petaFLOP/s
  - – ASC applications require >2 million messages/s and 1 GB memory per MPI task
  - – MPI tasks directly to cores, then the resulting interconnect reqs are not achievable
  - – MPI only leads to platform memory reqs that are not affordable or practical
  - – Divide and conquer by putting SMP parallelism within MPI tasks

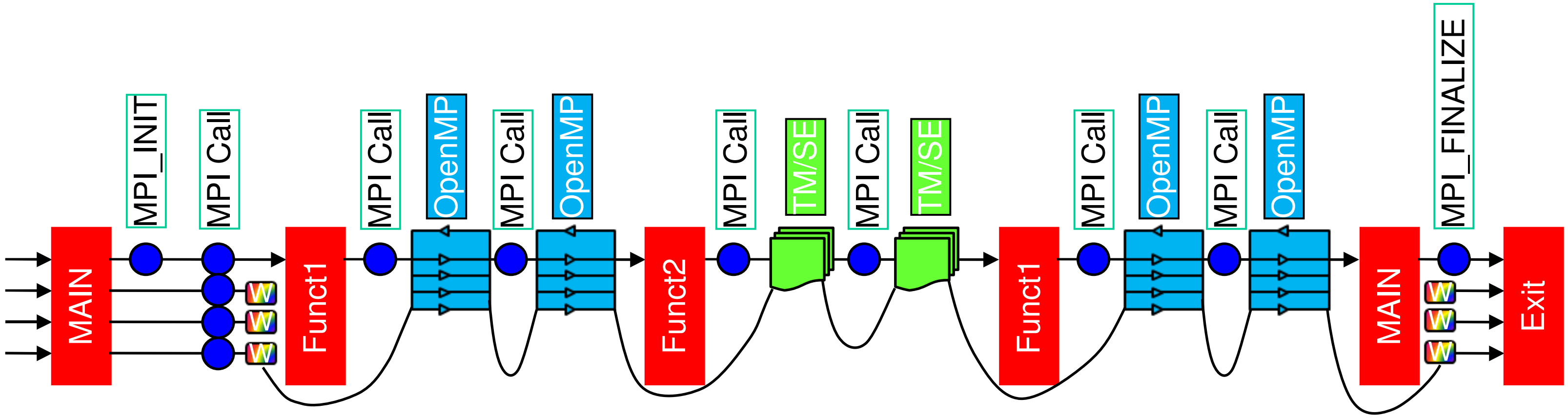
System	Cores	Nodes	MPI/node	SMP/MPI	Total MPI	Ratio:BGL
BGL	2	65,536	2	1	131,072	1.000
BGL	2	65,536	1	1-2	65,536	0.500
BGP(½)	4	36,864	4	1	147,456	1.125
LowPower	64	25,600	64	1	1,638,400	12.500
LowPower	64	25,600	4	16	102,400	0.781
LowPower	64	25,600	1	64	25,600	0.195
IntelTS	80	4,608	320	1	1,474,560	11.250
IntelTS	80	4,608	40	8	184,320	1.406
IntelTS	80	4,608	4	80	18,432	0.078

Parallelism of ¾ BGL for MPI with 16-64 SMP is more tractable!

Application Programming Model Requirements

- MPI Parallelism at top level
  - Static allocation of MPI tasks to nodes and sets of cores+threads
- Effectively absorb multiple cores+threads in MPI task
- Support multiple languages:
  - C, C++, Fortran03, and Python
- Allow different physics packages to express node concurrency in different ways

Unified Nested Node Concurrency



1) Pthreads born with MAIN
2) Only Thread0 calls functions to nest parallelism
3) Pthreads based MAIN calls OpenMP based Funct1
4) OpenMP Funct1 calls TM/SE based Funct2
5) Funct2 returns to OpenMP based Funct1
6) Funct1 returns to Pthreads based MAIN

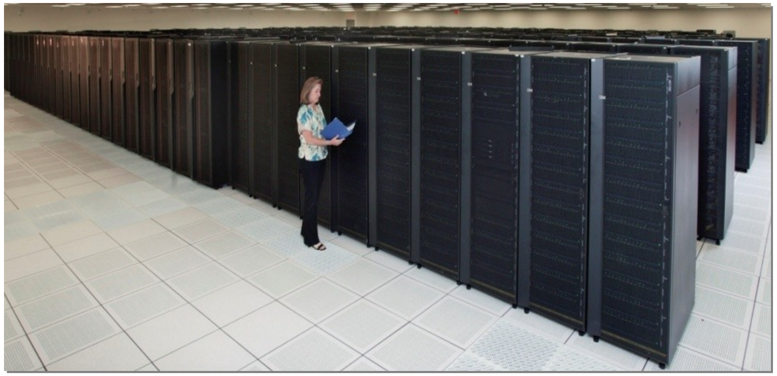
- MPI Tasks on a node are processes (one shown) with multiple OS threads (Thread0-3 shown)
- Thread0 is “Main thread” Thread1-3 are helper threads that morph from

Bronis de Supinski will talk more about the Sequoia software environment in the next talk

and OpenMP loops and locks

Sequoia Platform Target Performance is a Combination of Peak and Application Sustained Performance

- "Peak" of the machine is absolute maximum performance
  - FLOP/s = FLoating point OPeration per second
- Sustained is weighted average of five "marquee" benchmark code "Figure of Merit"
  - Four IDC package benchmarks and one "science workload" benchmark from SNL
  - FOM chosen to mimic "grind times" and factor out scaling issues



Purple – 0.1PF/s

BlueGene/L – 0.4 PF/s

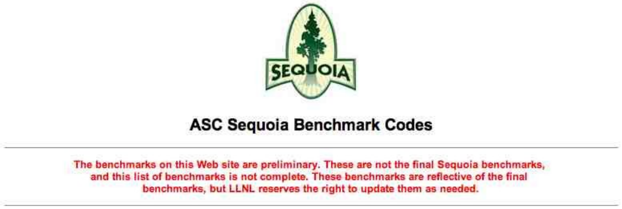
Sequoia Benchmarks have already incentivized the industry to work on problems relevant to our mission needs

- Production load “surrogates”
  - AMG – linear system solvers; MPI+OMP
  - IRS – Diffusion equation; MPI+OMP
  - SPhot – MC transport; MPI+OMP
  - UMT – Det. transport; FP, MPI+OMP, Python
- Science load “surrogate”
  - LAMMPS – Classical MD
- Functionality & Performance Tests
  - CLOMP – Threading overheads
  - Pynamic – Python; dyn. lib. perf.
  - MPI – Messaging performance
  - FTQ – OS Noise
  - IOR – IO performance
- Micro-kernels
  - Crystal_mk – SIMD
  - IRS_mk – SIMD
  - UMT_mk – Threading, FP perf.
  - AMG_mk – Threading, FP perf.
  - SPhot_mk – Integer per., branching

What’s missing?

- Hydrodynamics
- Structural mechanics
- Quantum MD

System	OS	CPU	Compiler
Purple	AIX	Power5	XLC
BG/L	LWK	PPC440	XLC
Atlas	Linux	Opteron	PathScale
Red Storm	LWK	Opteron	PGI



ASC Sequoia Benchmark Codes

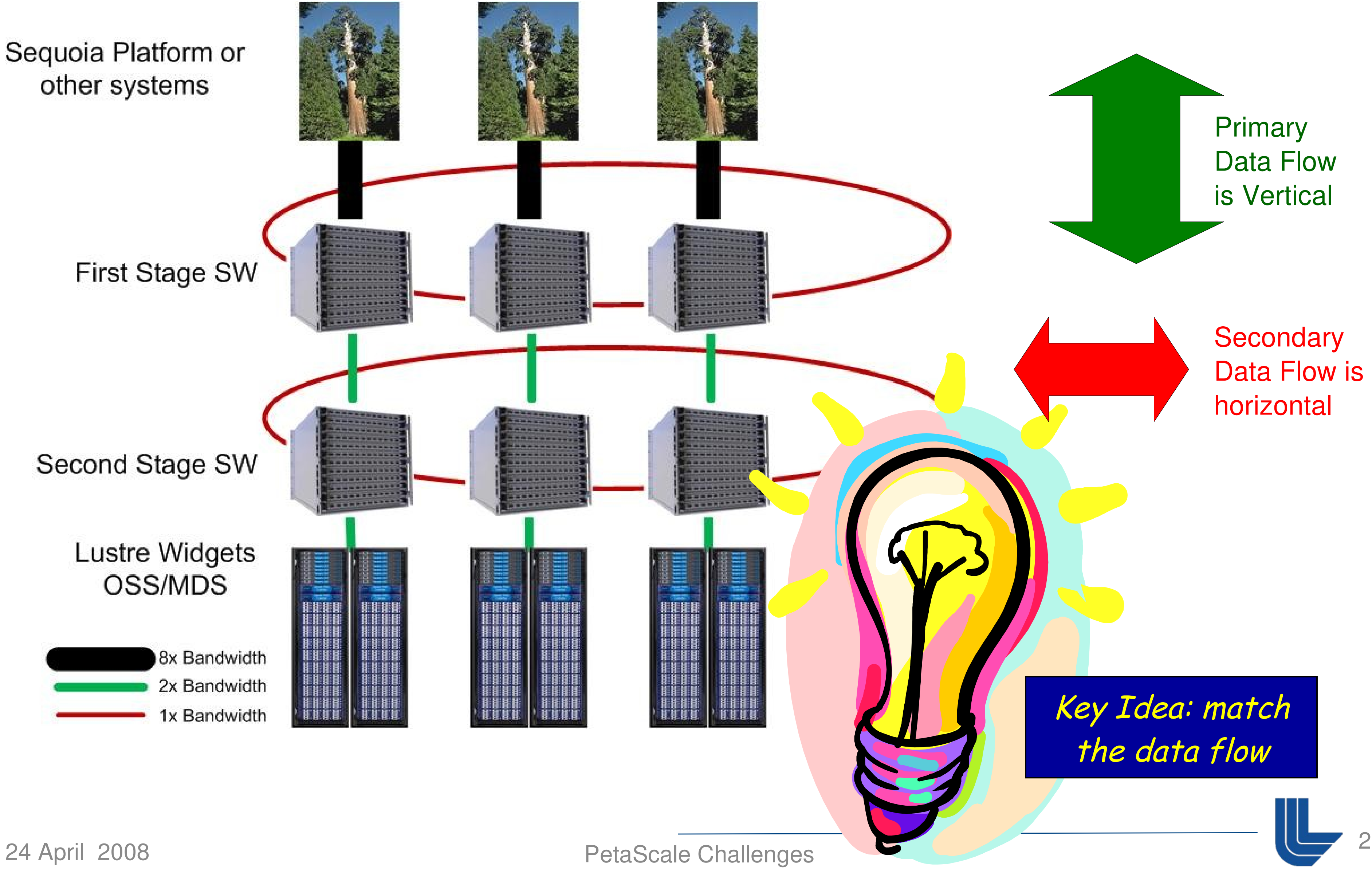
The benchmarks on this Web site are preliminary. These are not the final Sequoia benchmarks, and this list of benchmarks is not complete. These benchmarks are reflective of the final benchmarks, but LLNL reserves the right to update them as needed.

https://asc.llnl.gov/sequoia/benchmarks/

Sequoia SAN Challenges

- 100K-1M MPI tasks → ~1K I/O nodes → 100's storage widgets
- Turns lots of sequential I/O streams into many random I/O streams
- More intelligence in the network HW (NIC & switches) for dealing with congestion
- Requires an end-to-end SAN solution that is lacking most offerings
- Will Ethernet (40/100GbE) meet our cost, performance, and feature requirements?
- Can InfiniBand build on its success as a cluster interconnect and move out into the SAN?
- How do we design a cost effective, scalable, expandable SAN?

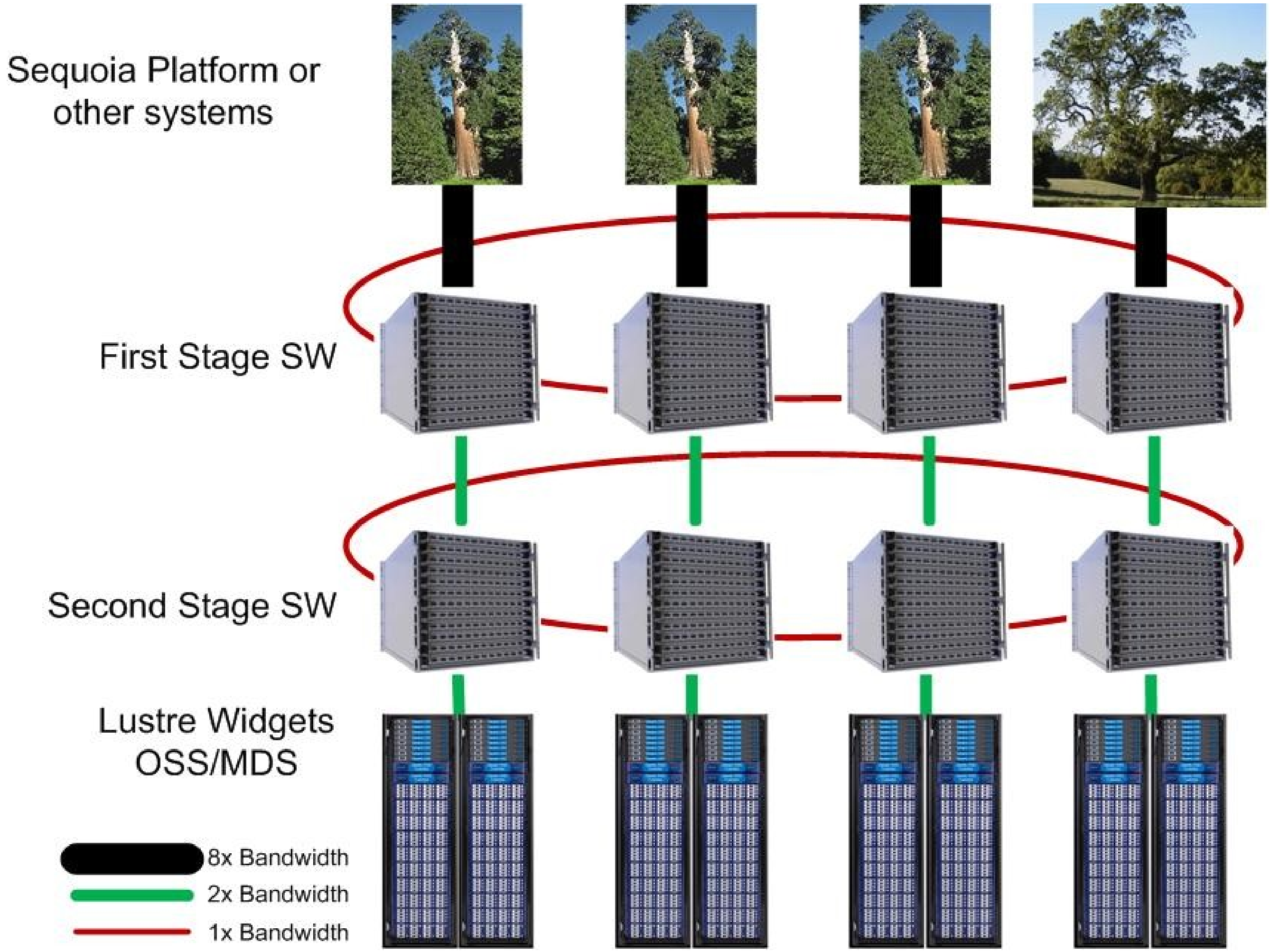
Sequoia Target SAN Environment is based on 2D Torus+Mesh topology with carefully balanced bandwidths

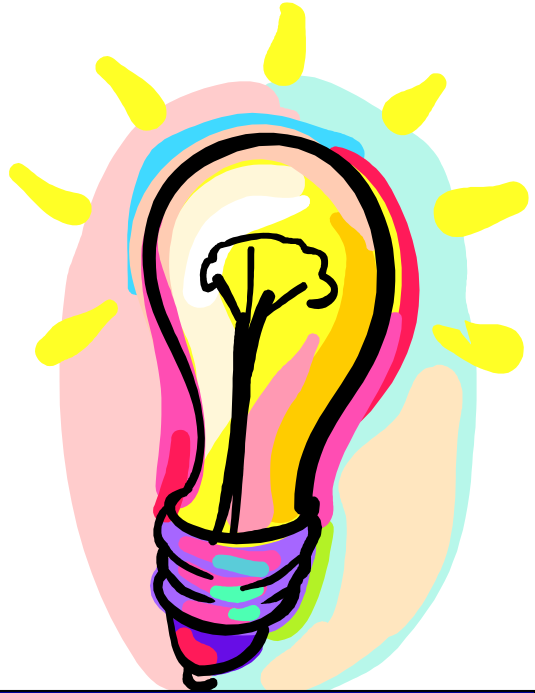


Sequoia Platform or other systems
First Stage SW
Second Stage SW
Lustre Widgets OSS/MDS
8x Bandwidth
2x Bandwidth
1x Bandwidth
Primary Data Flow is Vertical
Secondary Data Flow is horizontal

Key Idea: match the data flow

Additional System, SAN and Lustre Widget resources can be added incrementally without disrupting production





Key Idea: Allow for incremental expansion

Sequoia Procurement Timeline

- Platform
  - CD0 and CD1 signed
  - Gathering final technical input for SOW now
  - Should have RFP package together in a month
  - Two month DOE contracts review
    - Release DRAFT to industry while under DOE procurement review
  - Two week response period
  - Evaluation and selection process similar to Purple
  - 1-3 month subcontract negotiation endurance contest
  - Two month DOE contracts review
- SAN & Lustre Widgets
  - Market survey for last 18 months
  - Strategy agreed upon, SOW under construction
  - RFPs finalized after Platform DRAFT on the street

Summary

- Significant challenges remain in delivering PetaScale predictive simulations to problems of national benefit
- Sequoia is a carefully choreographed risk mitigation strategy to develop and deliver a huge leap forward in computing power to the National Stockpile Stewardship Program
- Sequoia will work for weapons science and integrated design codes when delivered because of our evolutionary approach to yield a revolutionary advance on multiple fronts
- This represents major innovations in procurement strategies, technology developments and collaborations by ASC to enable the petascale era

Sequoia will be the engine for the ASC's move towards more predictive simulation in the next decade

scaLE the Science, not just the codes